DOCUMENT RESUME

ED 161 943 TM 007 978

AUTHOR Petrosko, Joseph M.
TITLE Evolution of Educational Measurement in the 1970's: Changes in Elementary Level Standardized Tests.
PUB DATE Mar 78
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978); For related documents, see ED 044 446 and 143 670

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Affective Tests; Cognitive Tests; Educational Practice; *Educational Trends; Elementary Education; *Evaluation Criteria; Evaluation Methods; *Standardized Tests; *Student Testing; Test Reliability; *Test Reviews; Test Validity

ABSTRACT

Test evaluation summaries completed by the Center for the Study of Evaluation in 1970 and 1976 were used to determine changes in test quantity and quality among elementary-level standardized instruments. In the earlier studies, the instruments were rated in four general areas: measurement validity, examinee appropriateness, administrative usability, and normed technical excellence. Ratings covered critical indicators of test quality including reported validity and reliability and quality of score distribution. To determine changes in quantity, a cross-tabulation was constructed for 1970 and 1976 data showing the number of tests available for 41 educational goal areas at each grade level. The qualitative analysis focused on tests of: attitudes, values and motivation; reasoning, arithmetic operations, and reading readiness. Quality ratings were compared for concurrent and predictive validity and test reliability. The number of tests evaluated in each educational goal area in 1970 and 1976 are included, as well as the numbers and percentages of tests rated for validity and reliability. Results indicated that the quantity of elementary level standardized tests increased greatly and that increases were proportionally consistent within subject areas. However, despite an enormous growth in the number of tests, a parallel growth in quality did not occur. (Author/JAC)

Evolution of Educational Measurement in the 1970's: Changes in Elementary Level Standardized Tests

Joseph M. Petrosko
University of Louisville

Paper presented at the Annual Meeting of the National Council on Measurement in Education
Toronto, March 1978
Printed in U.S.A.

There have been many changes in educational measurement in the last several years. Changes have been evident in such things as (a) the measurement philosophy embodied in instruments, (b) the educational topics covered by instruments, and (c) the efforts made to deal with problems of test bias. The testing field has been broadened by the expansion of interest in criterion-referenced measurement. Theoretical concepts that were a gleam in a researcher's eye a decade ago have become reality in the form of commercially available tests. New subjects in the curriculum have led to newly developed tests. For example, emerging areas of education (e.g., occupational and career education, moral education) have matured to such a degree that educators are now interested in assessing educational outcomes in these areas.

Few would deny that it is a valuable enterprise to trace the evolution of educational measurement. But exactly how this should be done is another matter. The subject can be approached from many different angles. The present study approached the problem by dealing with two major issues. First, attention was focused on determining the areas of the curriculum that have seen an increase in testing options. In other words, what changes have occurred in the quantity of available tests and what curricular areas have been affected? Secondly, the study dealt with the issue of test quality: Irrespective of changes in the number of tests, has the measurement sophistication of instruments changed?

It is impossible in a single study to cover the entire measurement field. The study was limited to standardized tests aimed at elementary school students (i.e., grades 1-6). To examine changes in tests, a unique data base was used: test evaluation summaries completed by the Center for the Study of Evaluation. These test evaluations resulted from a large-scale

project involving a quality assessment of all standardized tests available in the United States. In the course of the project, tests were categorized by grade level and educational goal area, and then tests were evaluated for psychometric quality on over 20 criteria of excellence.

Using data assembled at the beginning of the decade (Hoepfner, Strickland, Stangel, Jansen, & Patalino, 1970) and data from the middle of the decade (Hoepfner, Bastone, Ogilvie, Hunter, Sparta, Grothe, Shani, Hufano, Goldstein, Williams, & Smith, 1976) a comparison was made of changes in test quantity and test quality among elementary-level standarized instruments.

Method

Procedure

The procedure for acquiring, categorizing and rating tests was the same in 1970 and 1976. First, all commercially available standardized tests at the elementary education level were ordered from publishers. Then the tests (including the subtests they contained) were categorized by grade level and by educational goal area (e.g., curriculum topics such as mathematics or reading). In 1970, the grade levels were 1, 3, 5 and 6; in 1976, the levels were grades 1, 2, 3-4, 5-6. For both sets of ratings, 41 educational goal areas were used. These covered the entire range of educational topics in the cognitive, affective and psychomotor domains. Based on the content of its items (e.g., arithmetic, reading, social studies) an instrument was categorized into a particular goal area.

After being categorized, the test was evaluated for quality. Each instrument was rated on more than 20 criteria of excellence. The criteria were grouped into four general areas: measurement validity, examinee appropriateness, administrative usability and normed technical excellence.

Ratings covered critical indicators of test quality, for example, reported validity and reliability and quality of score distribution.

Ratings were independently performed by two trained test reviewers; [illegible]ases of disagreement between the two, a third rater adjudicated disa[illegible]ents. All raters had the same information about each test--a standard s[illegible]en set consisting of the test itself and, in most cases, a technical manual or other type of supporting information. In assessing validity and reliability, only publisher-supplied information was used. No attempt was made to search through the research literature and find studies that employed a particular instrument.

Of necessity, the preceding description of test evaluation procedures has been brief. Complete details are available in Hoepfner et al. (1970) and Hoepfner et al. (1976).

Analysis

A straightforward analysis procedure was used to examine changes in tests. To determine changes in the number of instruments, a crosstabulation was constructed showing the number of tests available for 41 educational goal areas at each separate grade level. This was done both for 1970 and 1976 data. In the analysis of test quality, the study concentrated on tests in several important educational areas: (a) tests of attitudes, values and motivation, (b) tests of reasoning, (c) tests of arithmetic operations (i.e., computational ability) and, (d) tests of reading readiness. In each of these four areas, quality ratings were compared for two sets of evaluation criteria: (a) concurrent and predictive validity, and (b) test reliability. For each criterion, the number and percentage of tests at each level of quality were recorded.

Some thought was given to using inferential statistics to test

hypotheses regarding differences in test quality between 1970 and 1976. It was finally decided to forego such analyses since the data used in the study can be reasonably assumed to represent populations of tests rather than samples from populations. Inferential tests were, therefore, not reported.

Results

Changes in the Quantity of Tests

In the first part of the analysis, the numbers of tests in the various educational goal categories were compared for 1970 and 1976. It was discovered that there was a substantial increase in the number of instruments. In 1970, a total of 1,686 test evaluations were completed; in 1976, the number had risen to 9,127. For both occasions when evaluations were performed, more tests were found at the higher rather than the lower grade levels. The largest numbers of tests were located in areas of the curriculum that might be termed the "3 R's." Educational goal areas involving reading, writing, and arithmetic had a large number of tests. In addition to these, important segments of the cognitive and affective educational domains showed extensive coverage--personal temperament (e.g., tests of emotional stability), attitudes, values and motivation (e.g., tests of self-esteem and attitude toward school), and reasoning (e.g., tests of intelligence).

Table 1 shows the number of tests evaluated in 1970, Table 2 gives similar information for 1976. It should be noted that the grade level categories differed somewhat for the two sets of data. Also, the educational goal categories were different, but only slightly. Additions and deletions were made so that the 1976 goal list reflected an up-to-date picture of educational offerings at the elementary education level. For example, goal category 7 in 1976, Career Values and Understanding, illustrated a new

emphasis on career education that now extends down to the elementary level.

The data revealed a substantial increase, in absolute numbers, among almost all goal categories. Some educational areas were not well represented by instruments on both rating occasions. There was, relatively speaking, a small number of tests in arts and crafts, foreign language education, music, science, and social studies.

Table 1

Number of Elementary Standardized Tests Evaluated in 1970

Educational Goal Area	Grade 1	3	5	6
1. Temperament-Personal	13	14	17	24
2. Temperament-Social	15	17	20	27
3. Attitudes	4	5	5	7
4. Needs and Interests	0	0	5	23
5. Valuing Arts and Crafts	0	4	6	6
6. Producing Arts and Crafts	1	2	2	2
7. Understanding Arts and Crafts	0	3	3	4
8. Reasoning	53	43	50	47
9. Creativity	10	9	10	9
10. Memory	11	9	9	10
11. Foreign Language Skills	0	0	2	2
12. Foreign Language Assimilation	0	0	0	0
13. Language Construction	11	34	42	42
14. Reference Skills	0	4	13	14
15. Arithmetic Concepts	12	26	19	19
16. Arithmetic Operations	9	21	34	38
17. Mathematical Applications	5	8	12	15
18. Geometry	0	0	0	0
19. Measurement	0	1	2	2
20. Music Appreciation and Interest	0	0	0	0
21. Music Performance	0	0	1	1
22. Music Understanding	0	0	21	21

Table 1 (continued)

Educational Goal Area	Grade 1	3	5	6	
23. Health and Safety	1	3	8	8	
24. Physical Skills	17	11	4	4	
25. Sportsmanship	0	0	0	0	
26. Physical Education	0	0	0	0	
27. Oral-Aural Skills	10	3	2	2	
28. Word Recognition	46	45	32	23	
29. Reading Mechanics	14	17	16	13	
30. Reading Comprehension	84	97	88	91	
31. Reading Interpretation	0	2	11	13	
32. Reading Appreciation and Response	0	0	1	1	
33. Religious Knowledge	0	0	0	0	
34. Religious Belief	0	0	0	0	
35. Scientific Processes	0	0	2	2	
36. Scientific Knowledge	0	1	8	8	
37. Scientific Approach	0	0	0	0	
38. History and Civics	0	1	14	15	
39. Geography	1	2	11	11	
40. Sociology	0	0	1	1	
41. Application of Social Studies	0	1	5	5	
Totals	317	383	476	510	1686

Table 2

Number of Elementary Standardized Tests Evaluated in 1976

Educational Goal Area	Grade 1	2	3-4	5-6
1. Personal Temperament	174	171	231	220
2. Socialization	138	137	196	181
3. Attitudes, Values and Motivation	131	126	206	156
4. Valuing Art	15	16	20	19
5. Producing Art	9	8	8	8
6. Understanding Art	0	1	0	1
7. Career Values and Understanding	25	26	68	74
8. Understanding and Reasoning	253	210	304	295
9. Creativity and Judgment	28	29	34	36
10. Memory	85	75	90	75
11. Foreign Language Skills	11	26	28	92
12. Valuing a Foreign Language and Culture	2	3	5	5
13. Writing Skills	55	91	176	170
14. Reference and Study Skills	3	7	23	26
15. Understanding Math Concepts	46	34	55	38
16. Performing Arithmetic Operations	33	70	151	180
17. Applying and Valuing Mathematics	15	24	56	55
18. Geometry and Measurement Skills	6	6	24	29
19. Valuing Music	0	1	9	10
20. Performing in Music and Dance	0	0	0	0
21. Understanding Music	0	0	52	63
22. Sensory Perception	184	124	138	119

Table 2 (continued)

Educational Goal Area	Grade 1	2	3-4	5-6	
23. Psychomotor Skills	192	162	171	141	
24. Sports Skills	7	13	14	22	
25. Valuing Physical Education	1	0	1	2	
26. Health Habits and Understanding	10	10	18	17	
27. Understanding Hazards and Diseases	0	0	0	0	
28. Reading Readiness Skills	315	246	227	146	
29. Familiarity with Literature	0	0	1	1	
30. Reading with Understanding	115	182	296	226	
31. Reading Interpretation and Criticism	47	51	63	55	
32. Valuing Literature and Language	1	2	5	6	
33. Understanding Religion	0	0	0	2	
34. Personal Ethics and Religious Belief	11	10	12	9	
35. Investigating the Environment	4	4	2	3	
36. Understanding Science	3	3	11	14	
37. Valuing and Applying Science	1	2	6	5	
38. Understanding History and Civics	0	1	4	9	
39. Understanding Georgraphy	5	6	18	20	
40. Understanding Social Relationships	0	0	2	2	
41. Valuing and Applying Social Studies	13	14	21	20	
Totals	1938	1891	2746	2552	9127

Changes in the Quality of Tests

To approach the question of test quality, it was necessary to examine ratings of instruments on the various test evaluation criteria. There were over 20 criteria employed in evaluating tests (24 in 1970, 36 in 1976) and there were thousands of tests evaluated, so some simplification was required to avoid "data overload." It was decided to concentrate on tests in several key areas in the affective and cognitive educational domains. Each time, the same procedure was used. Test ratings were compared for concurrent and predictive validity (combined) and for three types of test reliability--test-retest, internal consistency and alternate form reliability.

In order to make test evaluations comparable for the 1970 and 1976 data, some adjustments were made in the ratings. In 1970, concurrent and predictive validity ratings were made using a 0 to 5 scale (ranging from "no evidence reported" to "exhaustive evidence"). In 1976, there were separate criteria for concurrent and predictive validity, each having a 0 to 2 scale (the higher a test was rated, the better its quality on each criterion). To facilitate comparison, the separate ratings in 1976 were added for each test to yield a new combined validity scale ranging in value from 0 to 4. The validity criterion designed for the present study contained 4 categories: high, medium, low, and very low or unreported. These reflected the following respective quality point designations: high--1970, 4 or 5 points, 1976, 4 points; medium--1970 and 1976, 3 points; low--1970 and 1976, 2 points; very low or unreported--1970 and 1976, 1 or 0 points. With reliability ratings, very minor changes were made to make the 1970 and 1976 data comparable. No new scales were constructed.

In making the ratings for validity and reliability, test evaluators searched through publisher-supplied information to arrive at a judgment.

Those tests that cited validity and reliability studies with high correlation coefficients were given the highest ratings. Medium correlations (ranging from .70 to .90) yielded medium ratings. If no studies were reported or if correlations were less than .70, the test was rated low.

In the first area studied, test evaluations were compared at the upper grade levels in an important part of the affective domain--the area of attitudes, values and motivation (1970, goal 3, Attitudes plus goal 4, Needs and Interests; 1976, goal 3, Attitudes, Values, and Motivation). This educational area covered such topics as attitude toward school, self-esteem, and achievement motivation. Table 3 reveals that the majority of such tests were rated low in validity and reliability in both 1970 and 1976. A few instruments received high ratings in reliability.

The area of reasoning was a part of the cognitive domain that had many tests in the category (1970, goal 8, Reasoning; 1976, goal 8, Understanding and Reasoning). The area covered instruments that measured skills traditionally included in intelligence tests--mental abilities such as classification, comprehension of information, logical reasoning and spatial reasoning. It was determined that there were increases in test quality among instruments in this category, at least in terms of slight increases in the absolute number of high quality instruments available. Table 4 shows that in 1976, there were more grade 5 and 6 tests with solid validity and reliability data than were available in 1970. The greatest increase in the number of high quality instruments occurred on the criterion of internal consistency reliability. The number of instruments with reported coefficients greater than .90 more than doubled.

The foregoing summaries concerned areas of great interest to educators, but not usually amenable to direct educational intervention. However, most

of the educational goal categories aimed at traditional curricular areas (i.e., school subjects). For example, several goal areas covered mathematics skills, the latter being a significant part of every elementary school curriculum. Table 5 gives ratings (at grades 5 and 6) of tests of arithmetic operations (1970, goal 16, Arithmetic Operations; 1976, goal 16, Performing Arithmetic Operations). These categories contained tests of ability to perform basic arithmetic: computation with whole numbers, fractions, decimals and percentages. Few tests at either year were high in validity or in test-retest and alternate form reliability. A slight increase did occur for instruments with high internal consistency coefficients.

The final educational category that was compared for changes in test quality was the area of reading readiness skills. Here, tests at grade 1 were examined. This age level was chosen since accurate information about reading readiness is most useful at the earliest primary school level. The subskills involved here included listening and speaking ability and word attack skills such as phonetic recognition (1970, goal 27, Oral-Aural Skills plus goal 28, Word Recognition; 1976, goal 28, Reading Readiness Skills). Table 6 reveals that not many changes occurred over the six year interlude. With this educational area, predictive validity is crucial to test usability, but ratings were low on the combined concurrent and predictive validity criterion. As with the other goal categories studied, there was a slight positive change in internal consistency reliability.

Table 3

Numbers and Percentages of Tests Rated for Validity and Reliability

Tests of Attitudes, Values and Motivation, Grades 5, 6

Evaluation Criterion	Year			
	1970		1976	
	n	%	n	%
Concurrent and predictive validity				
High	0	0	0	0
Medium	1	3	0	0
Low	1	3	0	0
Very low or unreported	38	94	156	100
Test reliability				
Test-retest coefficient				
$r > .90$	1	2	0	0
$.70 \leq r \leq .90$	12	30	4	3
$r < .70$ or unreported	27	68	152	97
Internal consistency coefficient				
$r > .90$	0	0	1	1
$.70 \leq r \leq .90$	14	35	3	2
$r < .70$ or unreported	26	65	152	97
Alternate form coefficient				
$r > .90$	0	0	0	0
$.70 \leq r \leq .90$	12	30	0	0
$r < .70$ or unreported	28	70	156	100

Table 4

Numbers and Percentages of Tests Rated for Validity and Reliability

Tests of Reasoning, Grades 5, 6

Evaluation Criterion	Year			
	1970		1976	
	n	%	n	%
Concurrent and predictive validity				
High	9	9	16	5
Medium	18	19	2	1
Low	6	6	52	18
Very low or unreported	64	66	225	76
Test reliability				
Test-retest coefficient				
$r > .90$	6	6	7	2
$.70 \leq r \leq .90$	23	24	28	9
$r < .70$ or unreported	68	70	260	89
Internal consistency coefficient				
$r > .90$	20	21	55	19
$.70 \leq r \leq .90$	34	35	47	16
$r < .70$ or unreported	43	44	193	65
Alternate form coefficient				
$r > .90$	5	5	13	4
$.70 \leq r \leq .90$	12	12	19	6
$r < .70$ or unreported	80	82	263	90

Table 5

Numbers and Percentages of Tests Rated for Validity and Reliability

Tests of Arithmetic Operations, Grades 5, 6

Evaluation Criterion	Year 1970 n	1970 %	1976 n	1976 %
Concurrent and predictive validity				
High	0	0	3	2
Medium	16	22	0	0
Low	2	3	19	11
Very low or unreported	54	75	158	87
Test reliability				
Test-retest coefficient				
$r > .90$	2	3	1	< 1
$.70 \leq r \leq .90$	11	15	6	3
$r < .70$ or unreported	59	82	173	97
Internal consistency coefficient				
$r > .90$	22	31	37	21
$.70 \leq r \leq .90$	23	31	20	11
$r < .70$ or unreported	27	38	123	68
Alternate form coefficient				
$r > .90$	0	0	0	0
$.70 \leq r \leq .90$	13	18	14	8
$r < .70$ or unreported	59	82	166	92

Table 6

Numbers and Percentages of Tests Rated for Validity and Reliability

Tests of Reading Readiness, Grade 1

Evaluation Criterion	Year 1970 n	1970 %	1976 n	1976 %
Concurrent and predictive validity				
High	2	4	3	1
Medium	3	5	2	1
Low	8	14	10	3
Very low or unreported	43	77	300	95
Test reliability				
Test-retest coefficient				
$r > .90$	1	2	1	< 1
$.70 \leq r \leq .90$	0	0	3	1
$r < .70$ or unreported	55	98	311	99
Internal consistency coefficient				
$r > .90$	13	23	24	8
$.70 \leq r \leq .90$	8	14	21	7
$r < .70$ or unreported	35	63	270	85
Alternate form coefficient				
$r > .90$	0	0	0	0
$.70 \leq r \leq .90$	3	5	2	1
$r < .70$ or unreported	53	95	313	99

Discussion

The results showed that the quantity of elementary level standardized tests increased greatly during the 1970's. Almost every major educational area had marked expansion in the availability of published instruments. Increases have been proportional. The areas of education well covered by tests in 1970 remained well covered in 1976. Unfortunately, the areas poorly covered at the beginning of the decade remained poorly covered at mid-decade. Areas such as arts and crafts, foreign language education, music, science and social studies are curricula without the measurement options they deserve. The major reason for this would probably derive from the non-traditional or heterogenous character of these subjects. Some school districts do not emphasize these subjects. When the subjects are taught, different content areas are emphasized in different districts. Without a uniform approach to subject-matter, test publishers are hard-pressed to develop tests that can be relevant to a variety of educational approaches.

The results regarding test quality ratings were depressing, if not altogether surprising. It would appear that, despite the enormous growth in the number of tests, a parallel growth in quality has not occurred. Tests of attitudes, reasoning, mathematics, and reading readiness have not shown noteworthy growth in quality. Of the test categories examined in this study, the most significant positive changes occurred among tests of reasoning, an area that encompassed IQ tests. It is possible that the tremendous public interest in intelligence tests in the early 1970's (especially regarding the testing of minorities) may have spurred test developers to refine intelligence measures to the greatest possible extent. Another reason for the rise in quality in this area may relate to its long history--it was one of the first areas addressed by test makers. There is a vast literature of

published research and test construction techniques from which authors of new tests can benefit.

There were several limitations to this study. First, it was necessary to construct a new scale to make predictive and concurrent validity ratings comparable for the two sets of ratings. This may have acted to "penalize" one set of ratings (such an occurrence is unlikely, but possible). A second limitation concerns differences in test evaluation procedures used in 1970 and 1976. Rating criteria were better defined and more stringently applied in 1976. This had the effect of requiring very convincing empirical evidence in order for a test to be rated high in validity or reliability. As a result, test ratings for 1976 may have been somewhat higher had some of the 1970 procedures been used in 1976. (Of course, the opposite is also true--had the 1976 procedures been applied to the 1970 data, the latter ratings would have been lower than they were.)

Despite these limitations, there is no reason to believe that the general findings were substantially affected. Regardless of minor differences in procedure, the conclusions remain--there has been an increase in the quantity of elementary level standardized tests and a negligible increase in quality.

The reader should note that some, if only a few, high-quality tests exist. The hundreds of mediocre instruments on the market should not obscure the good tests that are available. This study threw together highly developed instruments with some very poor specimens and the reader should not lose the proper perspective.

Perhaps the greatest value of the study was to reinforce the importance of each test consumer carefully considering the quality of a test before it is purchased. As a rule, poor tests outnumber good tests. This is true

regardless of the grade level or educational area of the test. For example, a study of secondary level tests (Petrosko, in press) has revealed findings basically congruent with the present study. An astute user of tests should consider relevant research and technical supporting information before making an expensive committment to purchase a particular measurement device.

References

Hoepfner, R., Bastone, M., Ogilvie, V. N., Hunter, R., Sparta, S., Grothe, C. R., Shani, E., Hufano, L., Goldstein, E., Williams, R. S., & Smith, K. O. CSE elementary school test evaluations. Los Angeles: Center for the Study of Evaluation, University of California, 1976.

Hoepfner, R., Strickland, G., Stangel, G., Jansen, P., & Patalino, M. CSE elementary school test evaluations. Los Angeles: Center for the Study of Evaluation, University of California, 1970.

Petrosko, J. M. The quality of standardized high school mathematics tests. Journal for Research in Mathematics Education, in press.

Acknowledgments

This study was based on data obtained from a project supported by the National Institute of Education (contract NE-C-00-3-0096). Conclusions do not necessarily reflect the views of that agency. The author wishes to thank Ralph Hoepfner of the System Development Corporation and Adrianne Bank and Russell Hunter of the Center for the Study of Evaluation.